Capellà i Ricart A, Roig Batalla A, 'The Expanding Notion of Special Categories of Data and its Limits' (2025) 2 EJPLT.
DOI: https://doi.org/10.57230/EJPLT252ACRARB

EJPLT
ISSN 2704-8012



The expanding notion of special categories of data and its limits

L'espansione della nozione di categorie speciali di dati e i loro limiti

ANNA CAPELLÀ I RICART
Postdoctoral Researcher
Universitat Autònoma de Barcelona

ANTONI ROIG BATALLA
Full professor of Constitutional Law
Universitat Autònoma de Barcelona

Abstract

Building on Solove's warnings on the uselessness of the notion of special categories of data, this article examines the expansion of the concept in the CJEU jurisprudence to assess if, indeed, it has become futile in the EU context. Our research explores how the concept has been expanded to inference, its specific application in the context of the right to be forgotten, and its role in social networks and competition law. We highlight that the current understanding of data of special categories presents limitations in terms of providing an effective protection in the field of big data analytics and we theorise on the suitability of shifting from an exhaustive list of data categories to a more general risk-based approach.

Muovendo dalle riflessioni critiche di Daniel J. Solove in merito alla scarsa utilità della nozione di categorie particolari di dati, il presente contributo esamina l'evoluzione ed espansione di tale concetto nella giurisprudenza della Corte di giustizia dell'Unione europea, al fine di verificare se esso sia divenuto, nel contesto dell'Unione europea, sostanzialmente privo di rilevanza operativa. L'analisi si concentra, in particolare, sull'estensione del concetto alle inferenze, sulla sua applicazione nel contesto del diritto all'oblio, nonché sul ruolo che esso assume nell'ambito dei social network e del diritto della concorrenza. Si evidenzia come l'attuale configurazione delle categorie particolari di dati presenti significative criticità sotto il profilo dell'effettività della tutela nell'era delle analisi di big data; in tale prospettiva, si propone una riflessione circa l'opportunità di superare un approccio fondato su un elenco tassativo di categorie, a favore di un modello più flessibile e generale, incentrato sulla valutazione del rischio.

© The author(s) 2025, published by Suor Orsola Benincasa Università Editrice.
This contribution is licensed under a Creative Common Attribution 4.0 International Licence CC-BY-NC-ND, all the details on the license are available at:
https://creativecommons.org/licenses/by-nc-nd/4.0/

Keywords: data of special categories; GDPR; inference; big data; expansion

Summary: 1. Introduction. – 2. The limited and selective expansion of the notion of data of special categories in the CJEU jurisprudence. – 2.1. The limited acceptance of inference in CJEU jurisprudence. – 2.2. The right to be forgotten: a latent generalisable expansion of special categories of data via the balancing test of Article 9(2)g) GDPR, and upon request. – 2.3. A potential game-changer: protecting data of special categories on social networks through competition law. – 3. The limits of the notion of data of special categories. – 3.1. Inferences of data of special categories in big data analytics. – 3.2. Special-category data, legitimate interest and AI model training. – 4. Conclusions.

1. Introduction.

Solove held recently that the General Data Protection Regulation (GDPR) and all laws concerning data of special categories 'make a fundamental conceptual mistake' when they focus on the nature of personal data: 'the borderlines of many categories are so blurry that they are useless'.[1] Instead, 'protections should be proportionate to the harm and risk' associated with the data processing. This Solove's article is intended as a contribution to the American debate over whether to adopt national privacy laws using an approach similar to that of the GDPR. However, it also takes a broader perspective – going beyond just the issue of special categories of data – and questions the very definition of personal data itself. In our opinion, this proposal also has value in contributing to the discussion on the jurisprudence of the Court of Justice of the European Union (CJEU).

The expansion of the notion of data of special categories has already been observed by many European scholars. Indeed, the CJEU has acknowledged some cases where special categories of data can be inferred, which, according to Solove, is one reason the boundaries between categories have become blurred.[2] Moreover, including data of special categories within the scope of the right to be forgotten has also contributed to the broadening of the notion. On the other hand, one of the most sensitive types of data – health data – has also drawn attention from a consumer protection perspective: as a result, even competing companies have sometimes been allowed to protect third-party' data of special categories. The final emerging trend for data of special categories originates from recent CJEU rulings involving social networks.

However, the expansion of sensitive data does not, *per se*, lead to Solove's conclusion that special categories are conceptually useless. For example,

[1] D Solove, 'Data is what data does. Regulating based on harm and risk instead of sensitive data' (2024) 118(4) Northwestern University Law Review 1081–1083. On the contrary, the Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR issued by the European Data Protection Board (Version 1.0, adopted on 8 October 2024) still focus on the nature of data, stating that 'the more sensitive or private the nature of the data to be processed, the more likely it is that the processing of such data will have a negative impact on the data subject'.

[2] ibid 1099.

enhanced protection for sensitive data remains valuable where such data is published in databases or newspapers.

That said, there may be a need for a complementary approach – one grounded in the assessment of risk or harm – beyond the literal interpretation of Article 9 of the GDPR. Considering the range of special categories listed in Article 9 and the even longer list of exceptions, this paper [merely] seeks to contribute to the broader, collective discussion on the future of special categories of data and their effective regulation.

For this purpose, we begin by analysing the CJEU cases in which the scope of special categories has been extended beyond the textual limits of Article 9 GDPR through inference, the right to be forgotten, and the consumer protection and social network context. Our aim is to determine whether the concept of special categories remains coherent, or whether it has, in practice, been gradually replaced by a broader risk-based approach (Section 2). If the latter is true, it would lend strong support to Solove's concerns. Then, we examine the limits of this heightened category of personal data (Section 3). In this regard, the existence of a high risk of discrimination and harm in big data analytics that current jurisprudence may not adequately address could be a strong reason to revisit the scope and content of Article 9 GDPR, or even to reconsider the continued usefulness of the notion of special categories of data altogether. Lastly, we will offer our conclusions (Section 4).

2. The limited and selective expansion of the notion of data of special categories in the CJEU jurisprudence.

In this section, we analyse how the CJEU has expanded the notion of sensitive data in three contexts: through inference (2.1), in relation to the right to be forgotten (2.2), and within the framework of competition law and social networks (2.3).

2.1. The limited acceptance of inference in CJEU jurisprudence.

Melissa Stock's comment on CJEU Case Vyriausioji tarnybinės etikos komisija (C-184/20)[3] clearly highlighted the expansion of the prohibition of processing of data of special categories to include inferred data.[4] In Lithuania, all public officers, as well as employees of non-governmental organisations, are required to declare their private data, including their name, personal identification number, social security number, income from self-employment, certain financial transactions, and any conflicts of interest. The same information must also be declared in relation to their spouse, cohabitee, or partner. Although most of this information is not published on the website of the Chief Official Ethics Commission, the name of the spouse, cohabitee, or partner is publicly available. As a result, the sexual orientation of the declarant and their partner

[3] Judgment of 1 August 2022, *Vyriausioji tarnybinės etikos komisija*, C-184/20, EU:C:2022:601.
[4] M Stock, 'Case Comment: The CJEU Expands the Scope of the Definition of Sensitive Personal Data in the GDPR: Case C-184/20 Vyriausioji tarnybinės etikos komisija' (2023) 24(1) Business Law International 93.

may be inferred from this data. The public disclosure of this information, though not being an explicit indication of a person's sexual orientation, was not strictly necessary and found to breach the principle of data minimisation. Moreover, publishing the name of a declarant's spouse, partner, or cohabitee exposes them to targeted advertising, commercial solicitation, or even criminal activity.

Therefore, the words 'revealing' and 'concerning', used to describe the prohibitions in Article 9(1) GDPR, [5] have been interpreted to include the inference of the information about sexual orientation from non-sensitive personal data – in this case, a partner's name: 'a wide interpretation of the terms 'special categories of personal data' and 'sensitive data' is confirmed by the objective of Directive 95/46 and the GDPR, (…) which is to ensure a high level of protection of the fundamental rights'.[6] Therefore, 'those provisions cannot be interpreted as meaning that the processing of personal data that are [leading] indirectly to reveal sensitive information concerning a natural person is excluded from the strengthened protection regime prescribed by those provisions'. [7] A broader perspective seems to emerge from this reasoning, particularly when considering how big data, machine learning (ML), or artificial intelligence (AI) may facilitate the inference of a 'hidden layer' of sensitive data and could blur the boundary between ordinary personal data and sensitive data,[8] as we will analyse in Section 3. These authors suggest that the nature of the processing, the objective suitability of the processed data for deriving the 'hidden layer', and the context-dependent sensitivity of the data should be taken into account.

However, the Court does not make any general statement regarding the broader applicability of this inference-based expansion, nor does it clearly define its consequences. Therefore, the Decision rather appears to represent an exception, limited in scope, and specifically applicable to non-public servants.[9] According to Hoppe, the different level of privacy protection in the CJEU Case Vyriausioji tarnybinės etikos komisija (C-184/20) as opposed to previous Decisions of the European Court of Human Rights (ECtHR), that were not even mentioned in the former Case, arises from the CJEU's finding that the

[5] Article 9.1 GDPR prohibits the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. There are some exceptions (Article 9(2)) that permit the processing of sensitive data: explicit consent (a); necessity for employment and social security purposes (b); protection of the vital interests of the data subject or of another natural person (c); legitimate activities by a foundation and similar bodies (d); related to personal data which are manifestly made public by the data subject (e); necessity for the establishment, exercise or defence of legal claims (f); necessity for reasons of substantial public interest, on the basis of union or Member State law (g); necessity for medicinal purposes or for the management of health systems and services (h); necessity for reasons of public interest in the area of public health (i); necessity for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes (j).

[6] Judgment of 1 August 2022, *Vyriausioji tarnybinės etikos komisija*, C-184/20, EU:C:2022:601, paragraph 125.

[7] ibid § 126.

[8] M Matejek and S Kremer, 'Sensible Daten soweit das Auge reicht: Der EuGH zapft den 'hidden layer' an' (2023) 4 Daten und Sicherheit 218.

[9] T Hoppe, 'Anticorruption, Privacy, and (Gay) Family – The CJEU Rules on Financial Disclosure of Public Officials (C-184/20)' (2013)17(4) ICL Journal 433–446.

declarant was not a public servant.[10] The Lithuanian law indeed treats the heads of NGOs receiving public funds as public servants, but the CJEU did not. This highlights, once again, the limited scope of this inference-based expansion: its applicability depends on the specific circumstances and is subject to the proportionality test (article 9(2) g) GDPR), on a case-by-case basis. If the country allows same-gender relationships, public officers could legitimately declare their partners to an agency monitoring asset declarations, but its public disclosure would require previous consent. In homophobic environments, the declaration and collection of this information from public authorities could contribute to discrimination on the ground of sexual orientation, no matter whether the declarant holds a public or a private position.[11]

2.2 The right to be forgotten: a latent generalisable expansion of special categories of data via the balancing test of Article 9(2)g) GDPR, and upon request.

The CJEU did not address the processing of sensitive personal data by search engine operators in its seminal Decision on the right to be forgotten, also known as the right to erasure or de-referencing, in Google Spain and Google (Case C-131/12).[12] The question was later answered in GC and Others (C-136/17), where the Court held there is no general prohibition on the processing of sensitive personal data by search engine operators.[13] Indeed, such operators may lawfully process sensitive data, even in the absence of a relevant exception under Article 9(2) GDPR, provided that the data subject does not object. Consequently, the level of protection afforded to sensitive data is effectively reduced unless the data subject actively asserts their rights.[14] Nonetheless, this creates a practical framework for search engine operators. The exceptions allowing them to process sensitive data could be: explicit consent (Art. 9(2) a) GDPR), which is rarely practical in this case; the data being manifestly made public by the data subject, though this exception was not cited in the Decision; and, almost notably, processing for reasons of substantial public interest (Art. 9(2) g) GDPR. Specifically, the Court mentions the right of freedom of information for Internet users (Art. 11 of the Charter of Fundamental Rights of the European Union).

The expansion of the notion of special categories of data in the context of search engine operators is, thus, based on a balancing of the right to data protection with the freedom of information, mainly, but not exclusively, in relation to journalistic content. Even where no explicit legal ground for personal data processing applies and the principles of Article 5 GDPR are not fully respected, the processing of sensitive data by search engines may still be

[10] Previous related Cases non mentioned in Case C-184/20 are *Wypych v Poland* (dec.), no. 2428/05, ECtHR 2005 and *Xhoxhaj v Albania*, no. 15227/19, ECtHR 2021.
[11] Hoppe, *Anticorruption, Privacy and (Gay) Family* (n 9).
[12] Judgment of 13 May 2014, *Google Spain and Google*, C-131/12, EU:C:2014:317.
[13] Judgment of 24 September 2019, *GC and Others,* C-136/17, EU:C:2019:773.
[14] J Globocnik 'The Right to Be Forgotten is Taking Shape: CJEU Judgments in GC and Others (C-136/17) and Google v CNIL (C-507/17)' (2020) 69(4) GRUR International 380–388.

deemed lawful if the balancing exercise favours it.[15] This assessment is explicitly tailored to search engine operators, but its rationale is potentially generalisable to other similar scenarios where there is no possibility to operate an *ex ante* and systematic control over the data to be processed, such as web scraping[16] or development and training of AI models.

The Court has therefore not established a blanket prohibition of the processing of sensitive personal data by search engine operators. Instead, they will have to monitor the processing *ex post* and decide, using a balancing of rights upon data subject's request and on a case-by-case basis, whether it is allowed or not to process the sensitive data. This approach may well extend to similar future cases, beyond search engines; though, its potential awaits further clarification in subsequent decisions of the CJEU.

2.3 A potential game-changer: protecting data of special categories on social networks through competition law.

The most recent expansive CJEU jurisprudence on special categories of data relates to social networks and competition law. Two leading cases need to be considered for this purpose: Meta Platforms and others (C-252/21) and Schrems (C-446/21).[17]

In the Case Meta Platforms and others, the CJEU held that a competition authority may consider data protection infringements when assessing abuse of a dominant position, pursuant to Article 102 TFUE, provided such consideration is necessary to establish the abuse. Competition authorities may require collaboration of national data protection authorities and must always follow their jurisprudence. As such, this opens the door to strengthening the protection of sensitive data when dealing with digital economy companies, such as social networks. Indeed, processing of personal data by an online social network may be prohibited when it reveals any of the sensitive categories listed in Article 9(1) GDPR, subject to derogations of Article 9(2) GDPR, irrespective of whether the information is accurate or whether the controller intends to obtain it.[18] Visiting websites or apps in certain cases may indirectly reveal such information without explicit disclosure by the data subject, thus constituting the processing of sensitive data.

The derogation under Article 9(2) e) GDPR requires an explicit and clear action by the data subject to make the data accessible to the public. When visiting websites or apps that may reveal sensitive information, users do not necessarily intend to make such information public; at most, such data might be intended for the website or app operator. Moreover, interactions via 'like' or 'share' buttons are governed by user-specific settings, and public disclosure to unlimited number of people requires explicit consent on specific individual setting in this sense. On the other hand, if the collected data set contains both

[15] Judgment of 24 September 2019, *GC and Others*, C-136/17, EU:C:2019:773, paragraph 75.
[16] Globocnik, *The Right to Be Forgotten* (n 14) 387.
[17] Judgment of 4 July 2023, *Meta Platform and others*, C-252/21, EU:C:2023:537.
[18] ibid §§ 66–69.

sensitive and non-sensitive data and no derogation under Article 9(2) GDPR applies, the general prohibition of Article 9(1) GDPR remains in effect.

Another example of the protection of special categories of data in social networks is the Case Schrems (C-446/21).[19] This Case is another opportunity for the Court to revisit its previous jurisprudence of Meta Platforms and others (C-252/21). Specifically, Article 9(2) e) GDPR is revisited as exception in the contexts of a user openly discussing their sexual orientation during a panel discussion. Although the data subject manifestly made this information public, the Court ruled that this alone does not permit the social network operator to process all other data related to that person's sexual orientation – such as data gathered from third-party websites and apps – for the purposes of delivering personalised advertisements.

In summary, this line of jurisprudence offers a more substantial expansion of the protection of special categories of data than previous developments. In sharp contrast with the Third-Party Doctrine in the USA, the CJEU does not treat social networks as public spaces, where users lose all expectation of privacy. Nor does it treat them as private spaces, where a reasonable expectation of data protection would generally exist. Rather, a nuanced approach is preferred and, in certain cases, specific and explicit consent is required for public disclosure, and the threshold for this consent remains high.

3. The limits of the notion of data of special categories.

While in the previous section we have analysed how the CJEU has extended the notion of data of special categories in various contexts, in this section, we explore the limitations that this notion presents as currently interpreted by the Court.

In the field of AI, it has been argued that the strengthened protection regime envisaged for data of special categories under the GDPR might be insufficient to address AI's discriminatory impacts or its implications for individuals' privacy and reputation. We analyse how this regime, as interpreted by the CJEU, may fail to provide an adequate protection standard in the context of big data analytics (3.1). Then, we study how, assuming that all personal data has the potential to become sensitive challenges the GDPR's protection of special categories of data limits the possibility of relying on legitimate interest as a legal basis for development and training of AI models (3.2).

3.1 Inferences of data of special categories in big data analytics.

It has been questioned whether the categorisation of sensitive data is useful in big data analytics. As Williams *et al.* explain, big data is composed of densely packed traces of individuals' lives pieced. Within such datasets, certain personal characteristics frequently appear together ('correlations' or 'patterns'). [20]

[19] Judgment of 4 October 2024, *Schrems*, C-446/21, EU:C:2024:834.
[20] BA Williams and others, 'How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications' (2018) 8 Journal of Information Policy 78–115.

These patterns may exist within a single individual's data (e.g. tennis class every Thursday), or across groups (e.g. residents of a certain neighbourhood tending to pay rent late). AI and ML systems can analyse these datasets, structure the data and create detailed individual profiles that necessarily include highly sensitive information. It is not necessary for sensitive data to be present *per se* in these datasets; such data may be inferred (directly or through the combination of personal or non-personal data).

The ECtHR Guide on its case law regarding data protection warns that there are categories of personal data other than those categorised as sensitive data that are also of concern 'especially with the increasingly sophisticated surveillance techniques and the capacity of information and communication technologies to make the daily life of data subjects more difficult'. [21] This remark questions the usefulness of any *numerus clausus* list of sensitive data categories. In this regard, it is unclear why the categories of data listed in Article 9 GDPR are considered special and other data, such as sex, social class, IP address or home address, is not.[22] The envisaged special categories do not align precisely with the protected grounds of the European Union (EU) antidiscrimination directives (sex, racial or ethnic origin, sexual orientation, religion or belief, age, or disability). The CJEU in the Case Meta Platforms and Others (C-252/21) specifies that data should be considered sensitive if such data allows for the inference of information falling within one of the categories listed in Article 9(1) GDPR. [23] Consequently, information revealing other characteristics (e.g. data revealing sex, economic or labour situation) that are not specified in Article 9 GDPR (but may equally have discriminatory impact or implications for privacy or reputation) falls outside the scope of special categories.

To this end, geo-location or browsing history can be used to draw precise conclusions about people's private live, such as daily routines or permanent or temporary residence. [24] Furthermore, nearly all apps make use of highly sensitive and personally identifiable information.[25] Such a massive volume of data increases risks of discrimination, exposure of sensitive information, and privacy violations. Within this context, the interplay between Article 9 GDPR and big data analytics becomes complex.

Quinn and Malgieri explain this challenge in the following terms: if only data listed in Article 9 GDPR is considered sensitive, other data can be used to deduce the same protected information. Conversely, if all data that can indirectly reveal sensitive information is considered sensitive, then nearly all

[21] ECtHR, Guide to the Case-Law of the European Court of Human Rights, Data protection, 31 December 2020, Updated on 31 August 2024. Available in <https://ks.echr.coe.int/documents/d/echr-ks/guide_data_protection_eng>, 15.
[22] Solove, 'Data is what data does.' (n 1) 1086.
[23] Judgment of 4 July 2023, *Meta Platform and others*, C-252/21, EU:C:2023:537, paragraph 68.
[24] JA Huerta Anguiano '"Naturaleza intrínseca", "contexto" o "finalidad" en la determinación del carácter sensible de los datos personales' (2020) 1(10) Estudios En Derecho a La Información 1–31.
[25] AS Devi and A Chinnasamy, 'Privacy Preservation of sensitive data in big data analytics – A survey' (2021) 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), 1–5.

personal data could fall within the scope of special categories of data, particularly given the combinatorial power of big data analytics.[26]

In the case Meta Platforms and Others (C-252/21), the CJEU stated that the processing of a dataset is prohibited if it is impossible to isolate non-sensitive data from sensitive data items at the time of collection. The Court refers to the occasions where the dataset contains at least one sensitive data item (see Section 2.3).

What does the CJEU mean when referring to 'sensitive data item'? As we referenced in Section 2.1, in the Case Vyriausioji tarnybinės etikos komisija (C-184/20), the Court asserts that information indirectly revealing sensitive information falls within the scope of Article 9 GDPR.[27] Advocate General Pikamäe affirms in his Opinion on the Case that the use of the verb 'reveal' includes an intellectual exercise involving deduction or cross-referencing.[28] Moreover, the Article 29 Working Parity (A29WP) further clarifies that the term 'data revealing' includes not only inherently sensitive data, but also such data from which sensitive information regarding an individual may be inferred.[29] As a result, it can be argued that inferred sensitive data, which is common in big data analytics,[30] should also receive heightened protection.

Nevertheless, as we previously pointed out, the Decision in Vyriausioji tarnybinės etikos komisija (C-184/20) seems exceptional, and later judgments indicate that the CJEU has not further extended this doctrine to big data analytics when stated the expansion of this principle regarding the inference in this Case. In this sense, it can be questioned whether the mechanism of 'separat[ing] the data items from each other at the time of collection'[31] referred in the Meta Platform and others (C-252/21) judgement is feasible in big data analytics. Van de Waerdt argues that this judgement reaffirms traditional principles of the data protection legislation in an unambiguous way and may have serious implications for digital market companies, which are likely to process (accidentally or not) special categories of data regarding their users given their widespread data collection.[32] Solove highlights that the idea that special categories of data can be separated is impractical in big data settings, where most of personal data is interconnected and may be inferred by algorithms.[33] Malgieri and Comandé argue that the identification of sensitive data can be difficult in the world of undisclosed algorithms and machine-generated inferences and predictions, where even the controllers

[26] P Quinn and G Malgieri, 'The Difficulty of Defining Sensitive Data–The Concept of Sensitive Data in the EU Data Protection Framework' (2021) 22(8) German Law Journal 1583.
[27] ibid § 127.
[28] Opinion of Advocate General Pikamäe in the CJEU case *Vyriausioji tarnybinės etikos komisija* (C-184/20), delivered on 9 December 2021, EU:C:2021:991, paragraph 85.
[29] Article 29 Data Protection Working Party, Advice Paper on Special Categories of Data ('Sensitive Data'), 4th April 2011, Ref. Ares ((2011)444105-20/04/2011).
[30] The European Data Protection Board Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024, does not analyse the processing of special categories of data in AI models (§ 17).
[31] Judgment of 4 July 2023, *Meta Platform and others*, C-252/21, EU:C:2023:537, paragraph 89.
[32] PJ van de Waerdt, 'Meta v Bundeskartellamt: Something Old, Something New' 8(3) European Papers: A journal on law and integration 1077.
[33] Solove, 'Data is what data does.' (n 1) 1083.

themselves may not be able to identify such data. [34] In particular, the identification of sensitive data can be difficult if the results given by the AI system are not human-readable.

In this regard, and taking into account that operating an *ex ante* and systematic control of what data will be processed seems to be difficult, it may be appropriate to extend the doctrine established by the CJEU Case GC and others (C-136/17) to big data analytics: data could be lawfully processed to develop and train AI models – even without an applicable exception under Article 9(2) GDPR – unless the data subject actively exercise their rights.

Alternative solutions have been proposed by the legal scholars to address the problem of the effectiveness of the GDPR's current model in light of the inference of sensitive data from other data. Nissenbaum suggests that categorising data as sensitive should be based on the context of the processing. [35] Purtova proposes that personal data should be protected according to potential individual or public negative consequences of the data processing. [36] Lastly, Quinn and Malgieri argue that data should only be considered sensitive if the 'intention of the data controller is to process or uncover sensitive information or if it is reasonably foreseeable that, in a given context, the data in question could be used to reveal or to infer sensitive aspects of data subjects'.[37]

In relation to this last proposal, the intention qualifier does not seem to be a suitable solution in big data scenarios, since it can be difficult or impossible to prove. As an example, in anti-discrimination law, there is a phenomenon called 'masking', that can be described as intentional discrimination disguised through technical opacity and complexity as unintentional discrimination.[38]

Furthermore, as we referenced in Section 2.3, the CJEU in the Case Meta Platforms and Others (C-252/21), states that it is irrelevant whether the information revealed by the processing operation in question is accurate, or whether the controller is acting with the aim of obtaining information that falls within one of the special categories.[39]

Fabris *et al.* argue that special categories of data are infrequent and difficult to process, hence limiting fairness approaches to proxy reduction or removal[40] (enabled by Article 10(5) AI Act).[41] The authors propose that the EU AI Act's

---

[34] G Malgieri and G Comandé, 'Sensitive-by-distance: quasi-health data in the algorithmic era' (2017) 26(3) Information and Communications Technology Law 229.

[35] H Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford University Press 2009).

[36] N Purtova, 'The law of everything: Broad concept of personal data and future of EU data protection law' (2018) 10(1) Law, Innovation and Technology 40. The author highlights that the rapid progress of (re)identification technologies and failure of anonymisation play an important role in the argument that everything will soon fall within the definition of personal data (any information relating to an identified or identifiable natural person) as interpreted by the A29WP and the CJEU.

[37] Quinn and Malgieri, *The Difficulty of Defining Sensitive Data* (n 26) 1612.

[38] JA Kroll and others, 'Accountable Algorithms' (2017) 165(3) U. Pa. L. Rev. 633–705.

[39] Judgment of 4 July 2023, *Meta Platform and others*, C-252/21, EU:C:2023:537, paragraph 69.

[40] A Fabris and others, 'Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey' (2025) 16 ACM Trans. Intell. Syst. Technol 1, 40.

[41] This Article provides another exception to allow the processing of sensitive data. The Article enables providers of high-risk AI systems to exceptionally process special categories of personal data when strictly necessary for the purpose of ensuring bias detection and correction, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons. Certain conditions for the processing must be met: the same purpose cannot be effectively fulfilled by processing other data (a); technical limitations are

exception to Article 9(1) GDPR be expanded to cover other scenarios where processing of special categories of data is necessary to prevent violations of certain rights of the data subjects (e.g. non-discrimination, privacy).

In summary, if the CJEU expands the notion of special categories of data to big data analytics, and a harm or risk-based approach is adopted, the concept of sensitive data may effectively collapse, considering that all personal and non-personal data in big data analytics may become sensitive information in a fraction of a second.[42] The absence of a clear regulatory framework for what constitutes sensitive data, especially in the context of big data analytics, can have further implications, as we address in the following section.

3.2 Special-category data, legitimate interest and AI model training.

An issue that arises from the ambiguous protection of inferred sensitive data in big data analytics is that, depending on what data is considered sensitive, the training and development of AI models may be either permitted or prohibited. The lack of guidance from the CJEU jurisprudence has resulted in limitations on the protection that can be derived from the concept of special categories of data.

AI systems, particularly those based on ML, require the use of data during the training phase, prior to being applied to new data in the operational phase. Based on this distinction, the Commission Nationale de l'Informatique et des Libertés (CNIL) issued a guideline on artificial intelligence,[43] stating that the processing of personal data for training purposes is subject to the GDPR if the training phase is separated from the operational implementation of the AI system and its sole purpose is to develop or improve the performance of the AI system.

The Belgian Data Protection Authority, in the Decision 46/2024 of 15th March,[44] also stated that it was necessary to distinguish between the stage of creating or training the AI model itself, and the operational stage, in which personalised discounts were offered through the digital applications that utilised the models created during the previous stage.[45] In that case, the claimant argued that the defendant (a bank) was using costumer transaction data to create and train AI models for offering personalised discounts for third-party products and services – thus processing data for a purpose different from the original one (executing and recording transactions).[46] The defendant

implemented in relation to the reuse of the data (b); measures are implemented to avoid misuse and ensure that only authorised persons have access to the data (c); data is not transmitted, transferred or accessed by other parties (d); data is deleted once the bias is corrected or the retention period has reached the end (e); and a record of the processing activities is performed, including why it was strictly necessary and why the purpose could not be achieved by processing other data.

[42] B van der Sloot, 'Regulating non-personal data in the age of Big Data', in M Tzanou (ed.), *Health Data Privacy under the GDPR: Big Data Challenges and Regulatory Responses* (Routledge 2020) 85.

[43] Commission Nationale de l'Informatique et des Libertés (CNIL), AI: ensuring GDPR compliance (21st September 2022). Available at: <https://www.cnil.fr/en/ai-ensuring-gdpr-compliance>.

[44] Autorité de protection des données, Décision quant au fond 46/2024 du 15 mars 2024, Numéro de dossier: DOS-2019-05837. Available in: <https://www.autoriteprotectiondonnees.be/publications/decision-quant-au-fond-n0-46-2024.pdf>.

[45] ibid § 45.

[46] ibid § 26.

argued that the legal basis for the processing of personal data at the training stage was legitimate interest (requiring that data subjects be informed and not object), and the legal basis for the operational stage was explicit consent (requiring that data subjects explicitly consent).[47] The Data Protection Authority accepted the defendant's argument regarding the distinction between two stages (training and operation) of the AI model and the use of legitimate interest as a legal basis for the training phase. Korff asserts, on the contrary, that the purpose of both stages was the same – direct marketing – and therefore the creation and training of the model (stage one) cannot be viewed as an independent activity,[48] since the two stages had the same purpose.[49]

Using this example as a reference, it is reasonable to question whether distinguishing the creation and training of AI models, on the one hand, and their operational use, on the other, has implications for the protection (and the usefulness of the classification) of special categories of personal data.

Under the GDPR, for data processing to be lawful, the processing must rely on one of the lawful grounds listed in Article 6 for non-sensitive personal data, and one of the additional grounds provided in Article 9 if special categories of data are involved.[50] The controller must satisfy the requirements of both GDPR provisions (Article 6 and Article 9) when it processes special categories of personal data.[51] Subsequently, processing of sensitive data is only allowed when one of the specific additional conditions set out in Article 9(2) GDPR applies.

The issue, therefore, remains the following: if the AI model is created and trained by processing non-sensitive personal data, legitimate interest may serve as an appropriate legal basis, provided that data subjects are informed and not object.[52] Alternatively, if sensitive personal data was used during the creation and training of the AI model, legitimate interest alone cannot be sufficient: an exception from Article 9(2) must apply, and data subjects must be informed.[53]

---

[47] ibid § 45.

[48] D Korff, 'Applying the EU General Data Protection Regulation (GDPR) to the use of personal data to train an AI system: A follow-up to my note on a deeply defective and flawed Belgian decision' (2024a). Blog [Data protection and digital competition].

[49] This case would be different for General Purpose AI Models (regulated in Chapter V of the AI Act) because the creating and training of the AI model can serve a *variety of purposes* both for direct use as well as for integration in other AI systems (Article 3.66 AI Act).

[50] European Union Agency for Fundamental Rights, Hand- book on European data protection law 2018 edition (2018), § 117. Available at: <https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law-2018-edition>.

[51] The European Data Protection Board Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024, does not analyse the processing of special categories of data in AI models (footnote 47).

[52] Korff argues that 'legitimate interest' can only be relied on if it is considered that there is sensitive data in the dataset when the sensitive data: a) relates to members and "'close contacts' of political, philosophical, religious or trade union bodies and are only processed by those bodies; b) has been 'manifestly' made public by the data subject; or c) is (strictly) necessary in a legal claims context. See: D Korff, 'The legal bases for processing of non-sensitive and sensitive data including the 'legitimate interest' legal basis & the use of personal data to 'train' an AI model' (2024b). Blog [Data protection and digital competition].

[53] Korff, 'Applying the EU General Data Protection Regulation (GDPR) to the use of personal data to train an AI system' (n 48).

Considering the previously mentioned legal doctrine that claims that all datasets used in big data analytics contain sensitive data, legitimate interest (Article 6.1.f) can only be invoked for the creation and training AI models if an exception of Article 9(2) also applies.[54] As the CJEU pointed out in Meta Platforms and Others (C-252/21), if the dataset [e.g. used to create and train AI models] contains 'at least one sensitive data item', and none of the exceptions in Article 9(2) GDPR apply, the processing is prohibited.[55]

Alternatively, Velli suggests that legitimate interest 'appears especially relevant in the context of AI models in which there is no direct relationship between the data subjects and the controller'.[56] Likewise, Trigo Kramcsák considers that the legitimate interests pursued by the data controller or a third party emerge as a lawful ground for building AI and ML training datasets.[57] Despite acknowledging that the use of sensitive data during the training of AI models can be challenging (for relying on legitimate interest), the author concludes that it remains a viable option, given that 'AI models trained on inaccurate or biased data can lead to harmful outcomes for individuals'.[58] Another possible option, as previously discussed, would be to extrapolate from the CJEU doctrine in GC and others (C-136/17) and rely on 'reasons of substantial public interest' as a legal basis (Article 9(2) g) GDPR). In any case, all these interpretations imply a weakening of the protection granted to sensitive data when used in AI model training.

4. Conclusions.

We have described a limited and selective expansion of the notion of special categories of data in the jurisprudence of the CJEU. Inferred data has been included (not universally, but in the specific case examined) within the scope of the prohibition of processing of special categories of data. This expansion has also been studied in relation to the right to be forgotten, though it remains constrained by the need to balance competing rights and, almost notably, is typically triggered by a request from the data subject. Lastly, the considerations related to social networks and competition law may also contribute to extending the protection of special categories of data in public spaces.

Solove's concerns appear, therefore, unjustified, at least under the current implementations of the AI Act and the GDPR. However, this could change with the rise of big data analytics. Without broader recognition of the inference in respect of sensitive data, we are still far from achieving a reasonable expectation of protection for special categories of data in public spaces.

Additionally, there are several other reasons to remain concerned about the future of the concept of special categories of data: the blurring line between

[54] ibid.
[55] Judgment of 4 July 2023, *Meta Platform and others*, C-252/21, EU:C:2023:537, paragraph 89.
[56] F Velli 'Processing Personal Data in the Context of AI Models: EDPB's Opinion 28/2024' (2024) 9(3) European Papers: A journal on law and integration 1497, 1498.
[57] P Trigo Kramcsák, 'Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets?' (2023) 48 CLSR 1, 11.
[58] ibid 11.

special categories of data and ordinary personal data, the reliance on the nature of data to determine whether it is considered sensitive, and the growing academic support for allowing the use of legitimate interest as a legal bases for training AI models.

However, shifting from fixed special categories of data towards a broader risk-based approach, as the AI Act proposes, may not be sufficient to adequately protect data users from discrimination and privacy harms, unless we recognise the existence of different kind of risks.